Exploiting Turkish Wikipedia as a Semantic Resource for Text Classification

Mitat Poyraz, Murat C. Ganiz, Selim Akyokuş, Burak Görener
Computer Engineering Dept.
Doğuş University
Acıbadem, Kadıköy, 34722, Istanbul, Turkey
{mpoyraz, mcganiz, sakyokus, bgorener}@dogus.edu.tr

Abstract—**Majority of the existing text classification algorithms are based on the "bag of words" (BOW) approach, in which the documents are represented as weighted occurrence frequencies of individual terms. However, semantic relations between terms are ignored in this representation. There are several studies which address this problem by integrating background knowledge such as WordNet, ODP or Wikipedia as a semantic source. However, vast majority of these studies are applied to English texts and to the date there are no similar studies on classification of Turkish documents. We empirically analyze the effect of using Turkish Wikipedia (Vikipedi) as a semantic resource in classification of Turkish documents. Our results demonstrate that performance of classification algorithms can be improved by exploiting Vikipedi concepts. Additionally, we show that Vikipedi concepts have surprisingly large coverage in our datasets which mostly consist of Turkish newspaper articles.**

Keywords-Textual Data Mining, Text Classification, Turkish Text Classification, Wikipedia, Vikipedi, Semantic Algorithms

I. INTRODUCTION

With the exponential growth of the web, developing more efficient text classification algorithms are gaining importance. However, the majority of the existing text classification algorithms have been based on the "bag of words" (BOW) approach, in which the documents are represented as weighted occurrence frequencies of individual terms (words). The BOW approach is effective when the category of a document can be easily determined by a few keywords [1]. However, its performance is limited to more challenging tasks because it only takes the word frequencies into consideration and ignores the semantic relationships between words that do not co-occur literally [2]. The semantic relationships among words can be used to improve the performance of text mining algorithms.

To deal with the inefficiencies of BOW approach, several studies have been done to integrate background knowledge such as WordNet [3], Open Directory Project (ODP) [4] or Wikipedia [5] to the existing systems. [6] use WordNet to improve text clustering whereas [7] use it to increase text categorization performance. [8, 1, 9] propose a method to integrate Wikipedia and ODP into text classification systems.

There are several important studies which use Wikipedia as external knowledge resource to enrich text classification. However, these studies are applied to English texts. In this study, we focus on Turkish Wikipedia (Vikipedi) [10] and use Vikipedi titles which we call concepts to enhance classification of Turkish texts. Our results show that, we can improve text classification accuracy by exploiting Vikipedi concepts. To the best of our knowledge this is the first study to integrate Vikipedi as background knowledge on classification of Turkish documents.

Rest of the paper is organized as follows: Background and related work is discussed in Section II. In Section III, the details of our methods are presented. Experimental setup and results are presented in section IV and V, respectively, followed by the conclusions and future work in section VI.

II. RELATED WORK

There are several studies which use external resources such as WordNet and ODP in order to overcome the shortcomings of traditional BOW approach. In [6], authors use WordNet to improve the performance of text document clustering. They integrate WordNet to text document representation by adding directly the words of the concepts which appear in the document and utilize this approach by deleting the words that do not appear in the WordNet itself. They cluster their representation using bi-secting k-Means algorithm on the Reuters dataset and show an improvement of 8.4% compared to the representation without WordNet. In a similar study, WordNet is used as a lexical resource to increase the text categorization performance [7]. In this approach, they find the synonyms of each category in the training set and compute a degree of semantic relatedness for each word in the synonyms. Then, they integrate this information with the training set and utilize Rocchio and Widrow-Hoff algorithms to test their approach. Their results show that, integrating WordNet increases the precision rate 20% on average for both algorithms on the Reuters- dataset.

In one of the premier studies in the context of text classification, Gabrilovich and Markovitch employ a feature generator to enrich BOW approach by using ODP [8]. Feature generator represents each ODP concept as a vector by using the hierarchical and textual information. Then, it finds the ODP concepts that exist in the given text and adds these concepts as features to traditional BOW representation. They use Support Vector Machine (SVM) algorithm on Reuters, 20 Newsgroups and Movies datasets and enhance the categorization performance by using these newly generated features. According to their results, they show improvement for all datasets with an improvement of up to 3.6% in terms of BEP (Precision-recall break-even point) for Movies dataset.

Recent work has shown that, due to its huge coverage and rich encyclopedic knowledge, Wikipedia can also be used as a semantic resource for text mining. In one of these studies, authors adopt semantic relatedness measures originally developed for WordNet and use it for Wikipedia [12]. In their approach, they retrieve the Wikipedia pages and the category information for given word pairs to compute the semantic relatedness between words. They perform several experiments on the WordSimilarity-353 Test Collection and show that using Wikipedia increases the performance of similarity calculations. In a similar study, authors propose a novel method which they called Explicit Semantic Analysis (ESA) [9, 1]. In their approach, they use a semantic interpreter to represent each text document as a weighted vector of Wikipedia concepts. Then they add these Wikipedia concepts to traditional BOW approach as new features. They use the WordSimilarity-353 collection and Australian Broadcasting Corporation's news mail service as datasets. Their results show that ESA with Wikipedia improves the correlation of computed semantic relatedness score with human from 0.6 to 0.72 for texts.

Wikipedia has considerable information about the relation between pages such as hyperlinks, hyponym, synonyms and category pages. There has been limited work using the link information in Wikipedia. The authors of [13] propose a new method called Wikipedia Link-based Measure (WLM) that uses only the hyperlinks between articles to extract semantic relatedness from Wikipedia. In their approach, given two words, they identify the related articles using anchor texts. Their results show that, WLM outperform the method applied in [12] by 0.19 and being outperformed by ESA method by 0.08 in terms of correlation with manually defined judgments. [14] use info box, categorical and actual link information that appear in an article. They build a graph consisting of articles as vertices and links between articles as edges. In order to find semantic relatedness between two texts, first they map these texts to graph and perform Personalized PageRank algorithm [15] to find stationary distributions for each text. Then, by comparing these two stationary distributions, they compute the semantic relatedness between texts. They achieve an improvement of 0.04 over baseline ESA method on the Lee dataset. In a similar study, authors also form a graph using Wikipedia articles and hyperlinks [16], but unlike in [14], they use cosine similarity between the representative vectors of the articles which they call lexical links. Their results show that, random walk with combining lexical links and hyperlinks outperform baseline reaching 90.8%.

While Wikipedia has been efficiently used in integrating semantic knowledge into text mining, it is not as structured as WordNet. Therefore, there have been studies to transform Wikipedia into a more structured thesaurus [17]. In this study, they search Wikipedia concepts in a given document. After finding candidate concepts, they add these and their related concepts together into the document. The related concepts are found using the synonymy or hyponymy or hierarchical relation or associative relation or combination of these. They use Reuters, Ohsumed and 20 Newsgroups as datasets and Support Vector Machine algorithm to classify documents. They use precision-recall break-even point (BEP) to measure the classifiers' performance. Their baseline approach, performing no text augmentation, reaches 0.865 macro-averaged BEP point on 20Newsgroups dataset. They show an improvement of 4.5% over baseline when both associative concepts and hyponyms are added to the text.

Although there are numerous studies using English Wikipedia in semantic analysis, there are limited numbers of studies using Turkish Wikipedia (Vikipedi) for text mining [11, 18, 19, and 20]. Among these, authors of [11] employ Vikipedi to discover missing links in a Vikipedi article. In another study [20], they build an automatic Turkish document summarization system. Authors of [18] integrate semantic information to Suffix Tree Clustering algorithm by using Vikipedi. In another study, knowledge based word sense disambiguation methods are compared to Turkish texts [19]. They use Turkish WordNet as a primary knowledge base and Vikipedi as enrichment resource for word sense disambiguation. However, none of these studies, which use Vikipedi, are about text classification.

III. APPROACH

Wikipedia is a free, collaborative, multilingual Internet encyclopedia which is growing exponentially. In Wikipedia, each article describes a topic and there is rich structural information between articles such as synonymy and hyperlinks. Turkish version of Wikipedia is called Vikipedi. In order to exploit Vikipedi as a semantic resource for text classification, we use dump of Vikipedi in xml format which can be downloaded from its official site. Our system is summarized in Figure 1 and explained in details in the following sections.

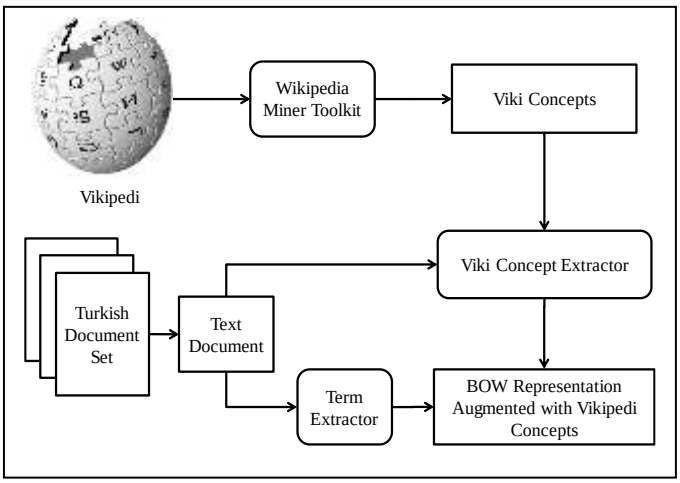

Figure 1. Design of the System

A. Wikipedia Miner Toolkit

Vikipedi dump dated September 8th, 2011 has 448,948 articles and title of each article describes a topic which we call concepts. These concepts can be single or multiple consecutive words that can be named entity, a compound word or a commonly used term in a specific domain. We obtain these concepts using Wikipedia Miner [21], which is a toolkit for discovering rich semantics encoded within Wikipedia (or Vikipedi). In general, Vikipedi has much larger coverage than manually built thesaurus like WordNet. However, it includes a

great deal of noise. In order to use Vikipedi more efficiently, we eliminate date concepts which include lists of events that occurred on a specific date. This preprocessing outputs 319,917 Vikipedi concepts, 219,755 of which consist of multiple words which we call Multiple Word Vikipedi Concepts (MWVC).

B. Term Extractor

Term Extractor produces a vector whose elements are term frequencies of words that appear in a given text document. Each text document is tokenized to words and these words are stemmed by Zemberek stemmer [22]. After stemming, stemmed terms are filtered by deleting stop words from documents. Finally, these terms are added to term-frequency vector with their frequencies. In Term Extractor, each text document is represented as term-frequency vector like traditional BOW approach; semantic relationships between terms are not captured in this representation.

C. Viki Concept Extractor

Viki Concept Extractor searches the Vikipedi concepts obtained in previous step in a given text document. These concepts may consist of a single or multiple consecutive words. Multiple word concepts are first tokenized to words and each word is stemmed by Zemberek stemmer. After stemming, each separated word is searched in the given text document. If all the separated words of a concept occur in text document consecutively, then this concept is added to term-frequency vector as a single entity. By this way, we add the semantic relationship between words which is ignored in BOW approach. In our study, the semantic relationship means, co-occurrence of multiple words literally. For example, "Mustafa Kemal Atatürk" is a single concept in Vikipedi and let's assume that these three words occur in a document consecutively. In BOW approach, all these words will be represented as separate terms in a vector space and their semantic relationship will be disregarded. On the other hand, in our model, Viki Concept Extractor will find that, these words occur consecutively in the document, so this concept is added to vector space as a single word. Thereby, we expand our document representation by augmenting new concepts obtained from Vikipedi.

IV. Experiment Setup

In order to analyze the performance of our model for text classification, we use four datasets consisting of Turkish newspaper articles namely 1150haber, AAHaber, AAHaber-18428 and Hürriyet-6c1k.

1150haber dataset consists of 1150 news articles evenly distributed in five categories [23]. These categories are ekonomi (economy), magazin (magazine), sağlık (health), politika (politics) and spor (sport). AAHaber-18428 is a dataset which consists of newspaper articles broadcasted by Turkish National News Agency, Anadolu Agency[1]. We compile 18,428 articles in seven categories namely dünya (world) , bilim-teknoloji (science-technology), ekonomi (economy), politika (politics), spor (sport), Türkiye (Turkey) and yaşam (life). Compared to the others, this dataset has a skewed class distribution. Majority of the documents are in politika (politics) and yaşam (life) categories. Another dataset which is also obtained from Anadolu Agency is called AAHaber [24]. AAHaber dataset has more articles (20,000) and these articles are evenly distributed in eight categories. These categories are culture art, economics, education science, environment health, politics, sports, Turkey, and world. Hürriyet-6c1k dataset is composed of 6,000 articles from Turkish newspaper Hürriyet [2] [25]. It contains 1,000 articles in each category: dünya (world), ekonomi (economy), gündem (agenda), siyaset (politics), spor (sport) and yaşam (life). Table 1 lists the number of categories (|C|) and the number of documents (|D|) in each dataset.

TABLE I. Description of the Datasets

Data Set	|C|	|D|
1150 HABER	5	1150
AAHABER	8	20000
AAHABER-18428	7	18428
HÜRRİYET-6C1K	6	6000

We employ multinomial Naïve Bayes (NB) and Support Vector Machine (SVM) [26] with linear kernel for our text classification experiments. Both SVM and NB are known to be among the best performing algorithms in text categorization. We use four different representations to analyze the effects of integrating Vikipedi as a semantic resource to text classification as follows:

- Traditional Bag of Words (BOW): Terms are extracted from a given document via Term Extractor module. Documents are represented by using the frequencies of these terms in the vector space.
- Bag of Words and Multiple Word Viki Concepts (BOW+MWVC): Vikipedi concepts which are composed of more than one word are identified. Traditional BOW representation of a document is augmented by adding these concepts as attributes.
- Multiple Word Viki Concepts (MWVC): Vikipedi concepts containing more than one word are extracted from a given document. Documents are represented by using only these multiple word Vikipedi concepts in the vector space.
- Viki Concepts (VC): All the Vikipedi concepts including one word concepts are extracted from a given document. Documents are represented by using only these single or multiple word Vikipedi concepts in the vector space.

Table 2 lists the number of terms in each of four different representations for each dataset. As it can be seen from the

1 http://www.aa.com.tr

2 http://ww.hurriyet.com.tr

table, Viki Concept Extractor expands BOW representation by adding considerable number of Vikipedi concepts. For example, BOW representation of Hürriyet-6c1k has 23,143 terms and 5,846 Vikipedi concepts are added to BOW+MWVC representation.

It is important to compare the number of terms in BOW and VC representations. For example, the number of terms in AAHaber in BOW is 19,739 whereas the number of terms in VC is 19,176. We speculate that Turkish Vikipedi includes the most of the terms in a given document; in this case the rate is 97.14%. If we compare the number of terms in other datasets, we see that Vikipedi covers 42.30%, 66.38%, 81.08% of terms on datasets 1150Haber, AAHaber-18428 and Hürriyet-6c1k respectively.

TABLE II. NUMBER OF TERMS IN EACH DATASET

DATA SET	BOW	BOW+ MWVC	MWVC	VC
1150 HABER	15280	16141	1141	6463
AAHABER	19739	26119	8291	19176
AAHABER-18428	48672	60945	15611	32310
HÜRRİYET-6C1K	23143	28989	7599	18765

V. EXPERIMENT RESULTS AND DISCUSSION

We apply Naïve Bayes Multinomial (NB) and Support Vector Machine (SVM) algorithms to each of our datasets. For each data set we performed 10-fold cross-validation and report average accuracy. The results of our experiments are given in Tables 3 to 6.

Table 3 shows accuracy of NB and SVM classifiers. NB and SVM columns list the accuracy of each classifier on baseline BOW representation. The other columns display, the accuracy results of classifiers on MWVC representation which is enriched by Vikipedi concepts. It is usually very difficult to get high performance increases in text classification when the classifier accuracies are high. When the classifier accuracy is comparatively low, like 78.5% in Hürriyet-6c1k dataset, higher increases in accuracy can be obtained (79.02%).

TABLE III. ACCURACY RESULTS FOR NB AND SVM

DATA SET	NB-BOW	NB-MWVC	SVM-BOW	SVM-MWVC
1150 HABER	93,13%	**93,22%**	87,39%	**87,92%**
AAHABER	81,65%	**82,02%**	80,24%	**80,30%**
AAHABER-18428	86,64%	**86,91%**	85,29%	**85,69%**
HÜRRİYET-6C1K	78,50%	**79,02%**	74,93%	**75,78%**

Vikipedi contains many different types of semantic relationships such as synonymy, polysemy, categorical information and hyperlinks between articles. In our study, we only use semantic relationship of words that co-occurs literally. As it can be seen in Table 4, we obtained a slight improvement in accuracy for all datasets as we expected. Even though the performance improvements are little, it is important to note that enriched representations increase the accuracy of the classifiers in all of the four datasets. This encourages us to further research on the use of Turkish Vikipedi as an external resource to improve text mining methods.

TABLE IV. ACCURACY IMPROVEMENT OVER BASELINE

DATA SET	NB	SVM
1150 HABER	0,093 %	0,597 %
AAHABER	0,447 %	0,075 %
AAHABER-18428	0,300 %	0,470 %
HÜRRİYET-6C1K	0,658 %	1,134 %

TABLE V. ACCURACY RESULTS AND DICTIONARY SIZE FOR NB

DATA SET	BOW	|V|	VC	|V|
1150 HABER	93,13±0,03	15280	91,39 ± 0,03	**6463**
AAHABER	81,65±0,01	19739	79,73±0,01	**19176**
AAHABER-18428	86,65±0,01	48672	83,89±0,01	**32310**
HÜRRİYET-6C1K	78,50±0,02	23143	75,72±0,03	**18765**

Table 5 shows the classifier accuracy of the NB algorithm on the four of our datasets and the number of terms in each dataset. The VC representation only uses the terms/concepts that appear in Vikipedi. As it can be seen from the table, there is not a great difference in the accuracy results between using BOW and VC representations in NB algorithm. However, in VC representation, the number of terms is relatively less than BOW representation. The performance of NB classifier does not reduce much despite the decreasing number of terms in VC representations. This shows that Vikipedi supplies well accepted and widely used concepts and these concepts have high discriminative power as features in classification.

TABLE VI. ACCURACY RESULTS AND DICTIONARY SIZE FOR SVM

DATA SET	BOW	|V|	VC	|V|
1150 HABER	87,39±0,03	15280	84,26 ± 0,04	**6463**
AAHABER	80,24±0,01	19739	79,40±0,01	**19176**
AAHABER-18428	85,29±0,01	48672	84,30±0,01	**32310**
HÜRRİYET6C1K	74,93±0,02	23143	73,93±0,01	**18765**

Table 6 shows the performance of SVM classifier on each of the four data sets. We observe similar results compared with NB classifier. Just as in NB, the accuracy results are very close

in BOW and VC representations when we apply SVM algorithm. Especially in Hürriyet6c1k dataset, the accuracy results are approximately the same even though VC representation includes 4,378 less terms than BOW representation.

VI. Conclusions and Future Work

Vikipedi is a massive resource of encyclopedic knowledge. There are several studies which use English Wikipedia as external knowledge resource to enrich text mining methods. However, there are a very limited number of studies using Turkish Wikipedia (Vikipedi). To the best of our knowledge, this is the first study that utilizes Turkish Vikipedi in order to enhance text classification. We extract Vikipedi concepts which are titles of Vikipedi articles and augment BOW representation by adding new concepts. Experimental results show that augmenting traditional BOW representation with Vikipedi Concepts improves the accuracy of classifiers. Another important finding is that Turkish Vikipedi covers the most of terms in our datasets which mainly consist of Turkish newspaper articles. For example, in AAHaber data set, Vikipedi includes 97.14% terms that appears among 20,000 documents. When we use only concepts included in Vikipedi for classification, experimental results show that classifier accuracies are not affected much. This shows that Vikipedi includes well accepted and widely used concepts and these concepts have high discriminative power as features in classification.

Vikipedi is a very rich information resource and contains semantic relationships such as synonymy, polysemy, hyponymy, associative and categorical information and hyperlinks between articles. In our study, we only used semantic relationship of words that co-occurs literally in titles of articles. As a future work, we plan to use other relationships that exist in Vikipedi to enhance text classification and exploit this rich information resource for improving other text mining methods.

Acknowledgment

This work is partially supported by TÜBİTAK under Grant No. 111E239.

References

[1] Gabrilovich, E. and Markovitch, S. (2006) Overcoming the brittleness bottleneck using Wikipedia: Enhancing text categorization with encyclopedic knowledge. In Proceedings of the Twenty-First National Conference on Artificial Intelligence , Boston, MA.

[2] Hu, J., Fang, L., Cao, Y., et al. Enhancing Text Clustering by Leveraging Wikipedia Semantics. In Proceedings of the 31st annual International ACM SIGIR Conference on Research and Development in Information Retrieval. (Singapore, July 20 – 24, 2008). ACM Press, New York, NY, 179-186.

[3] Miller, G. (1995). WordNet: A lexical database for English. CACM, 38, 39–41.

[4] ODP, Open Directory Project, http://dmoz.org

[5] Wikipedia, http://en.wikipedia.org/wiki/Wikipedia:About

[6] A. Hotho, S. Staab and G. Stumme. Wordnet improves text document clustering. In Proceedings of the Semantic Web Workshop at SIGIR'03

[7] Buenaga, M. de Buenaga Rodriguez, J. M. G. Hidalgo, and B. Dıaz-Agudo. Using WordNet to complement training information in text categorization. In Recent Advances in Natural Language Processing II , volume 189. 2000.

[8] Gabrilovich, E., and Markovitch, S. 2005. Feature generation for text categorization using world knowledge. IJCAI , 1048–1053.

[9] Gabrilovich, E. and Markovich, S. (2007) Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis. In Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI'07) , Hyderabad, India

[10] Vikipedi, http://tr.wikipedia.org/wiki/Ana_Sayfa

[11] O. Sunercan and A. Birturk, "Wikipedia Missing Link Discovery: A Comparative Study," in AAAI Spring Symposium on Linked Data Meets Artificial Intelligence (Linked AI 2010), ser. AAAI Spring Symposium, A. S. Symposium, Ed., Stanford, USA, 2010.

[12] Strube, M. and Ponzetto, S. P. (2006). WikiRelate! Computing Semantic Relatedness Using Wikipedia. In Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-06), pp.1419-1424.

[13] D. Milne and I.H. Witten. 2008. An effective, low-cost measure of semantic relatedness obtained from wikipedia links. In Proceedings of the first AAAI Workshop on Wikipedia and Artifical Intellegence (WIKIAI'08), Chicago, I.L.

[14] E. Yeh, D. Ramage, C. D. Manning, E. Agirre, and A.Soroa.2009. WikiWalk: Random walks on Wikipedia for semantic relatedness. In TextGraphs.

[15] T. H. Haveliwala. 2002. Topic-sensitive pagerank. In WWW '02, pages 517–526, New York, NY, USA ACM.

[16] Majid Yazdani, Andrei Popescu-Belis Using a Wikipedia-based semantic relatedness measure for document clustering Proceeding TextGraphs-6 Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing

[17] P. Wang, J. Hu, H. Zeng, and Z. Chen, "Using Wikipedia knowledge to improve text classification," Knowledge and Info. Systems, Springer-Verlag, Vol. 19, No. 3, 2009, pp. 265-281.

[18] C. Calli, G. Ucoluk, T. Sehitoglu. (2010) Improving Search Result Clustering By Integrating Semantic Information from Wikipedia, M.S Thesis

[19] A. Boynuegri, A. Birturk. (2010) Cross-Lingual Information Retrieval on Turkish and English Texts, M.S Thesis

[20] Aysun Güran and Nilgün Güler Bayazıt, "A New Preprocessing Phase for LSA-Based Turkish Text Summarization",Recent Advances in Computer Science and Information Engineering Lecture Notes in Electrical Engineering, 2012, Volume 124, 305-310, DOI: 10.1007/978-3-642-25781-0_46.

[21] D. Milne and I. H. Witten. An open-source toolkit for mining wikipedia. In Proc. New Zealand Computer Science Research Student Conf., volume 9, 2009.

[22] A.A. Akın , M.D. Akın, "Zemberek, an open source nlp framework for Turkic languages".

[23] M.F. Amasyalı, A. Beken, "Türkçe Kelimelerin Anlamsal Benzerliklerinin Ölçülmesi Ve Metin Sınıflandırmada Kullanılması", Siu 2009, Antalya.

[24] Tantug A.C., 2010. "Document Categorization with Modified Statistical Language Models for Agglutinative Languages", International Journal on Computational Intelligence Systems, Vol.:5 No:3 Models For Agglutinative Languages

[25] Torunoğlu, D., Çakırman, E., Ganiz, M.C., Akyokuş, S., Gürbüz, M.Z. (2011). Analysis of Preprocessing Methods on Classification of Turkish Texts. *INISTA 2011* , June, 2011, Istanbul, Turkiye.

[26] J. Platt: Fast Training Of Support Vector Machines Using Sequential Minimal Optimization. In B. Schoelkopf And C. Burges And A. Smola, Editors, Advances In Kernel Methods - Support Vector Learning, 1998